

GHAJAR EXHIBIT 49

2/26/2025

Richard Kadrey, et al. v. Meta Platforms, Inc.
Highly Confidential - Attorneys' Eyes Only

Lyle Ungar

Page 116

1 Q. 9-0. That's right. I'm sorry.

2 I realized I was reaching and
3 talking away from you as I was asking the
4 question.

5 So in paragraph 90 of your
6 opening report --

7 A. Give me a second.

8 Q. Sorry.

9 A. Sorry.

10 Q. I apologize.

11 A. All good.

12 Okay. I have found 90.

13 Q. Yes. So in paragraph 90 of your
14 opening report, you write that in order to
15 generate responses -- and here's where I'm
16 going to start reading [as read]:

17 "Models must be trained
18 on data that is sufficiently
19 large, diverse, and high
20 quality."

21 Right?

22 A. Correct.

23 Q. Okay. So when you are discussing
24 quality, you are not opining as to quality in
25 any economic sense, right?

2/26/2025

Richard Kadrey, et al. v. Meta Platforms, Inc.
Highly Confidential - Attorneys' Eyes Only

Lyle Ungar

Page 117

1 A. Correct.

2 Q. Meaning -- by "quality," you are not
3 providing any opinion about whether or not
4 "quality" means for something to be worth more
5 monetarily, right?

6 A. Correct.

7 Q. Okay. And just to be clear, you are
8 not offering any opinion about the economic
9 value as to any of the information that may
10 have been used as training data in this case,
11 correct?

12 A. Correct.

13 Q. Can you please define for me what you
14 mean by "quality" when you're using it in this
15 sentence?

16 A. Yes. Quality is a slightly broad
17 concept here.

18 But quality data, somewhat
19 tautologically, is data that leads to better
20 large language models. In particular, when we
21 train models, we like to throw out gibberish,
22 things that are non-human-readable.

23 The web is full of junk.

24 "Quality" often means that we want, for the
25 sort of applications that Meta or I care about,

2/26/2025

Richard Kadrey, et al. v. Meta Platforms, Inc.
Highly Confidential - Attorneys' Eyes Only

Lyle Ungar

Page 118

1 things that are not pornography.

2 So often there's a huge
3 filtering of removing pornography to try and
4 get high quality. Often try and remove
5 repetitions, duplication.

6 So good-quality data is
7 something that doesn't have the same thing over
8 and over and over again.

9 I could list ten other
10 attributes of quality.

11 Q. So would quality depend on -- well,
12 let's take a step back here. So I want to
13 tease out that answer a little bit.

14 A. Uh-hum.

15 Q. So you testified that "quality" means
16 that -- for the sort of applications that Meta
17 cares about, right -- I believe that's what you
18 said.

19 So would quality vary from LLM
20 designer to LLM designer depending on what that
21 person wants the LLM to do?

22 A. Absolutely. Some want to do things
23 that are good at writing software. Some want
24 them to be good at answering facts.

25 There's a different quality of

2/26/2025

Richard Kadrey, et al. v. Meta Platforms, Inc.
Highly Confidential - Attorneys' Eyes Only

Lyle Ungar

Page 119

1 data for doing software -- writing Python code
2 than there is for answering facts about world
3 quality -- world capitals.

4 Q. Okay. Could "quality" also mean, by
5 improving the performance of an LLM, certain
6 benchmarks?

7 A. Yes.

8 Q. And there are a number of
9 standardized benchmarks used by practitioners
10 in the generative AI field, correct?

11 A. There are.

12 Q. Okay. And if we turn to paragraph 96
13 of your opening report, you describe some of
14 the -- some pitfalls with data that might lead
15 to information not being high quality, right?

16 And I will turn your attention
17 to the first sentence on top of page 59 -- the
18 first full sentence of the top of page 59.

19 A. [As read]:

20 "Data with abundant
21 misspellings, toxic language,
22 or misinformation are
23 generally avoided."

24 Q. Yes.

25 A. Yes. Those are all indications of

2/26/2025

Richard Kadrey, et al. v. Meta Platforms, Inc.
Highly Confidential - Attorneys' Eyes Only

Lyle Ungar

Page 120

1 potential lower quality.

2 Q. Okay. And you are also -- I believe,
3 in your prior answer, you described for us an
4 example of pornography, right?

5 A. Correct. Again, depending upon the
6 application.

7 Q. So wouldn't it actually be the case
8 that quality -- so quality does -- so I believe
9 we discussed this briefly.

10 But quality is really dependent
11 on what you want your LLM to do, right?

12 ATTORNEY MORTON: Object to
13 form.

14 THE WITNESS: That's correct.

15 BY ATTORNEY YOUNG:

16 Q. Okay. So would quality, therefore,
17 be a somewhat subjective measure?

18 A. Yes. Although you could, if you had
19 a clear benchmark you're optimizing, put
20 something in and measure how much effect it has
21 on the benchmark.

22 Q. All right. For example, in your --
23 in your -- you conducted an experiment using
24 one of these benchmarks, right, MMLU?

25 A. I did.

2/26/2025

Richard Kadrey, et al. v. Meta Platforms, Inc.
Highly Confidential - Attorneys' Eyes Only

Lyle Ungar

Page 190

1 didn't -- really was just -- do you know why
2 LibGen was created, actually?

3 A. I do not know. I mean, many times
4 for things -- like, there's the -- GitHub
5 wasn't created as a training set either but is
6 widely used here.

7 So a bunch of things came from
8 different sources.

9 Q. By the way, do you know if GitHub is
10 one of the trainings that's in the Pile?

11 A. I don't know.

12 Q. Well -- okay. So back to -- back to
13 LibGen.

14 So -- so LibGen was not designed
15 as a training data set, right?

16 I believe that's where we left
17 off.

18 A. Correct.

19 Q. Okay. So in order to use LibGen
20 as -- if one were to use LibGen as training
21 data, it would then follow that one would
22 probably have to process the LibGen data?

23 A. All these data sets are always
24 processed. Pretty much nobody trains stuff off
25 the shelf because there's too much boilerplate

2/26/2025

Richard Kadrey, et al. v. Meta Platforms, Inc.
Highly Confidential - Attorneys' Eyes Only

Lyle Ungar

Page 191

1 background and headings. So we're always
2 stripping out the background information.

3 Q. Do you know if that's true for Books3
4 as well?

5 A. I would bet that most -- that most
6 people, when they do it, do an extra removal of
7 redundancy.

8 Q. Right. But I think there's a step
9 there.

10 Books3 is already processed,
11 correct, as training data?

12 A. I think maybe we're being confused.
13 There's several different kinds of processing
14 that can happen.

15 There's PDF-to-text processing.
16 There's removal-of-boilerplate processing.

17 So when you say "processed," I'm
18 not sure which one you're referring to.

19 Q. Okay. So Books3 was -- is a data set
20 that is designed for training generative AI
21 products, right?

22 ATTORNEY MORTON: Object to
23 form.

24 BY ATTORNEY YOUNG:

25 Q. Oh, excuse me.

2/26/2025

Richard Kadrey, et al. v. Meta Platforms, Inc.
Highly Confidential - Attorneys' Eyes Only

Lyle Ungar

Page 223

1 read, no. That sort of common knowledge
2 of movies is almost missing from Archive.

3 So Archive is mostly technical
4 papers, as you probably know. It's going
5 to give a very different set of things
6 from most of the MMLU-style questions of
7 the version you gave.

8 BY ATTORNEY YOUNG:

9 Q. Okay. But you expect, for example,
10 Archive may improve the performance of an LLM
11 to answer factual questions in certain domains
12 that may be contained in the Archive articles
13 that it's trained on, correct?

14 ATTORNEY MORTON: Object to
15 form.

16 THE WITNESS: Correct. More
17 technical, chemistry or physics or
18 computer science.

19 BY ATTORNEY YOUNG:

20 Q. So if you were to train an LLM on
21 certain textbooks, you would expect its ability
22 to answer factual questions about those domains
23 to increase, correct?

24 A. Correct.

25 Q. So, for example, if you were to train

2/26/2025

Richard Kadrey, et al. v. Meta Platforms, Inc.
Highly Confidential - Attorneys' Eyes Only

Lyle Ungar

Page 224

1 an LLM on a corpus of biology textbooks, you
2 would expect its knowledge of biology to
3 increase, right?

4 A. Yes. With the caveat you normally
5 want to throw up in a bunch of other texts too
6 to improve its strength. Agreed.

7 Q. Can you explain more about why you
8 would want to include other texts too? What do
9 you mean -- well, let me start...

10 What do you mean by "improve its
11 strength" there?

12 A. Improve the performance. So the goal
13 of these models is not just to answer biology
14 questions, and it's not -- certainly not to
15 store facts because it doesn't actually store
16 facts in any sort of a dictionary or database
17 or encyclopedia.

18 The goal is to be able to answer
19 questions, and so -- or hold conversations,
20 respond.

21 So, for example, having seen
22 more conversations or more text, sometimes even
23 seeing computer software helps them do better.
24 On other problems, it helps them reason better.

25 So, in general, having diversity

2/26/2025

Richard Kadrey, et al. v. Meta Platforms, Inc.
Highly Confidential - Attorneys' Eyes Only

Lyle Ungar

Page 225

1 is a good thing.

2 Q. So I understand.

3 So it's not enough for an LLM to
4 simply recall facts or store facts; it has to
5 be able to communicate semantically in a -- or
6 communicate in a semantic way, almost kind of
7 conversationally, right?

8 ATTORNEY MORTON: Object to
9 form.

10 THE WITNESS: Could be said more
11 precisely, but I think the gist is you're
12 moving in the right direction.

13 BY ATTORNEY YOUNG:

14 Q. So what would -- what sort of data
15 would you expect to improve an LLM's ability to
16 respond in that more conversational-type way?

17 ATTORNEY MORTON: Object to
18 form.

19 THE WITNESS: Well, two things.

20 Ideally, give it some
21 conversation data. The other piece is
22 that all of these models -- this is a
23 whole different direction and probably out
24 of scope -- are fine-tuned after the fact
25 where they say good response/bad response

2/26/2025

Richard Kadrey, et al. v. Meta Platforms, Inc.
Highly Confidential - Attorneys' Eyes Only

Lyle Ungar

Page 353

1 CERTIFICATE

2 I HEREBY CERTIFY that the
3 proceedings, evidence and objections are
4 contained fully and accurately in the
5 stenographic notes taken by me upon the
6 deposition of Lyle Ungar, taken on
7 2/26/25 and that this is a true and correct
8 transcript of same.

9

10

11

12

13

14

15

16

17

18

19

20

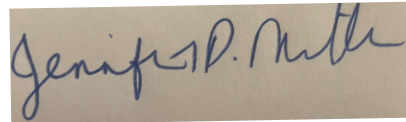
21

22

23

24

25

A handwritten signature in blue ink on a light brown rectangular piece of paper. The signature appears to read "Jennifer D. Miller".

Jennifer Miller, RMR, CCR, CRR
and Notary Public

(The foregoing certification of
this transcript does not apply to any
reproduction of the same by any means
unless under the direct control and/or
supervision of the certifying reporter.)